**贝叶斯推断及其互联网应用（二）：过滤垃圾邮件**

作者： [阮一峰](http://www.ruanyifeng.com/)

日期： [2011年8月27日](http://www.ruanyifeng.com/blog/2011/08/)

上一次，我介绍了贝叶斯推断的[原理](http://www.ruanyifeng.com/blog/2011/08/bayesian_inference_part_one.html)，今天讲如何将它用于垃圾邮件过滤。

========================================

**贝叶斯推断及其互联网应用**

作者：阮一峰



（接上文）

**七、什么是贝叶斯过滤器？**

垃圾邮件是一种令人头痛的顽症，困扰着所有的互联网用户。

正确识别垃圾邮件的技术难度非常大。传统的垃圾邮件过滤方法，主要有"关键词法"和"校验码法"等。前者的过滤依据是特定的词语；后者则是计算邮件文本的校验码，再与已知的垃圾邮件进行对比。它们的识别效果都不理想，而且很容易规避。

2002年，[Paul Graham](http://www.paulgraham.com/spam.html)提出使用"贝叶斯推断"过滤垃圾邮件。他说，这样做的效果，好得不可思议。1000封垃圾邮件可以过滤掉995封，且没有一个误判。

另外，这种过滤器还具有自我学习的功能，会根据新收到的邮件，不断调整。收到的垃圾邮件越多，它的准确率就越高。

**八、建立历史资料库**

贝叶斯过滤器是一种统计学过滤器，建立在已有的统计结果之上。所以，我们必须预先提供两组已经识别好的邮件，一组是正常邮件，另一组是垃圾邮件。

我们用这两组邮件，对过滤器进行"训练"。这两组邮件的规模越大，训练效果就越好。Paul Graham使用的邮件规模，是正常邮件和垃圾邮件各4000封。

"训练"过程很简单。首先，解析所有邮件，提取每一个词。然后，计算每个词语在正常邮件和垃圾邮件中的出现频率。比如，我们假定"sex"这个词，在4000封垃圾邮件中，有200封包含这个词，那么它的出现频率就是5%；而在4000封正常邮件中，只有2封包含这个词，那么出现频率就是0.05%。（【注释】如果某个词只出现在垃圾邮件中，Paul Graham就假定，它在正常邮件的出现频率是1%，反之亦然。这样做是为了避免概率为0。随着邮件数量的增加，计算结果会自动调整。）

有了这个初步的统计结果，过滤器就可以投入使用了。

**九、贝叶斯过滤器的使用过程**

现在，我们收到了一封新邮件。在未经统计分析之前，我们假定它是垃圾邮件的概率为50%。（【注释】有研究表明，用户收到的电子邮件中，80%是垃圾邮件。但是，这里仍然假定垃圾邮件的"先验概率"为50%。）

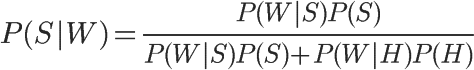
我们用S表示垃圾邮件（spam），H表示正常邮件（healthy）。因此，P(S)和P(H)的先验概率，都是50%。

http://chart.googleapis.com/chart?cht=tx&chl=P(S)%3DP(H)%3D50%25&chs=40

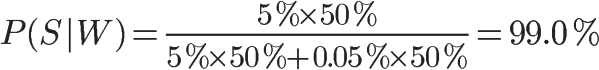
然后，对这封邮件进行解析，发现其中包含了sex这个词，请问这封邮件属于垃圾邮件的概率有多高？

我们用W表示"sex"这个词，那么问题就变成了如何计算P(S|W)的值，即在某个词语（W）已经存在的条件下，垃圾邮件（S）的概率有多大。

根据条件概率公式，马上可以写出



公式中，P(W|S)和P(W|H)的含义是，这个词语在垃圾邮件和正常邮件中，分别出现的概率。这两个值可以从历史资料库中得到，对sex这个词来说，上文假定它们分别等于5%和0.05%。另外，P(S)和P(H)的值，前面说过都等于50%。所以，马上可以计算P(S|W)的值：



因此，这封新邮件是垃圾邮件的概率等于99%。这说明，sex这个词的推断能力很强，将50%的"先验概率"一下子提高到了99%的"后验概率"。

**十、联合概率的计算**

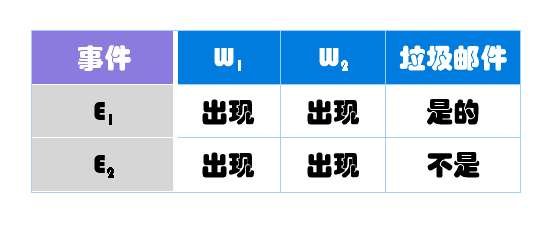
做完上面一步，请问我们能否得出结论，这封新邮件就是垃圾邮件？

回答是不能。因为一封邮件包含很多词语，一些词语（比如sex）说这是垃圾邮件，另一些说这不是。你怎么知道以哪个词为准？

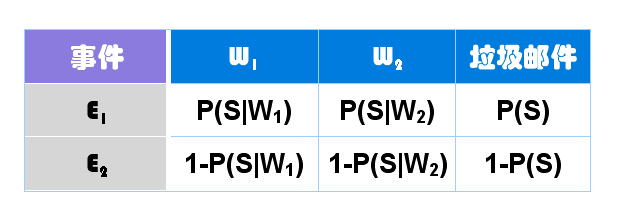
Paul Graham的做法是，选出这封信中P(S|W)最高的15个词，计算它们的联合概率。（【注释】如果有的词是第一次出现，无法计算P(S|W)，Paul Graham就假定这个值等于0.4。因为垃圾邮件用的往往都是某些固定的词语，所以如果你从来没见过某个词，它多半是一个正常的词。）

所谓联合概率，就是指在多个事件发生的情况下，另一个事件发生概率有多大。比如，已知W1和W2是两个不同的词语，它们都出现在某封电子邮件之中，那么这封邮件是垃圾邮件的概率，就是联合概率。

在已知W1和W2的情况下，无非就是两种结果：垃圾邮件（事件E1）或正常邮件（事件E2）。



其中，W1、W2和垃圾邮件的概率分别如下：

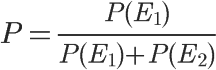


如果假定所有事件都是独立事件（【注释】严格地说，这个假定不成立，但是这里可以忽略），那么就可以计算P(E1)和P(E2)：

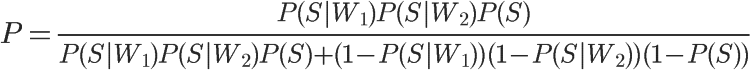
http://chart.googleapis.com/chart?cht=tx&chl=P(E_%7B1%7D)%3DP(S%7CW_%7B1%7D)P(S%7CW_%7B2%7D)P(S)&chs=40

http://chart.googleapis.com/chart?cht=tx&chl=P(E_%7B2%7D)%3D(1-P(S%7CW_%7B1%7D))(1-P(S%7CW_%7B2%7D))(1-P(S))&chs=40

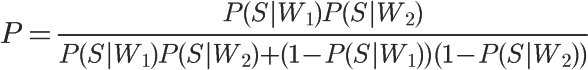
又由于在W1和W2已经发生的情况下，垃圾邮件的概率等于下面的式子：



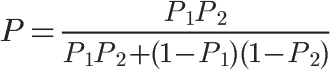
即



将P(S)等于0.5代入，得到



将P(S|W1)记为P1，P(S|W2)记为P2，公式就变成



这就是联合概率的计算公式。如果你不是很理解，点击[这里](http://www.mathpages.com/home/kmath267.htm)查看更多的解释。

**十一、最终的计算公式**

将上面的公式扩展到15个词的情况，就得到了最终的概率计算公式：

一封邮件是不是垃圾邮件，就用这个式子进行计算。这时我们还需要一个用于比较的门槛值。Paul Graham的门槛值是0.9，概率大于0.9，表示15个词联合认定，这封邮件有90%以上的可能属于垃圾邮件；概率小于0.9，就表示是正常邮件。

有了这个公式以后，一封正常的信件即使出现sex这个词，也不会被认定为垃圾邮件了。

（完）